A detailed analysis of kernel parameters in Gaussian process-based optimization

Hossein Mohammadi, Rodolphe Le Riche, Eric Touboul

A detailed analysis of kernel parameters in Gaussian process-based optimization

Hossein Mohammadi [*1, 2], Rodolphe Le Riche [†2, 1], and Eric Touboul [‡1, 2]

[1]Ecole des Mines de Saint Etienne, H. Fayol Institute
[2]CNRS LIMOS, UMR 5168

Abstract

The need for globally optimizing expensive-to-evaluate functions frequently occurs in many real-world applications. Among the methods developed for solving such problems, Efficient Global Optimization (EGO) is regarded as one of the state-of-the-art unconstrained continuous optimization algorithms. The most important control on the efficiency of EGO is the Gaussian process covariance function which must be chosen together with the objective function. Traditionally, a parameterized family of covariance functions is considered whose parameters are learned by maximum likelihood or cross-validation. In this report, we theoretically and empirically analyze the effect of length-scale covariance parameters and nugget on the design of experiments generated by EGO and the associated optimization performance.

keywords: Continuous global optimization; EGO; Gaussian processes

1 Introduction

We wish to find the global minimum of a function f, $\min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x})$, where the search space $\mathcal{S} = [LB, UB]^d$ is a compact subset of $\mathbb{R}^d$. We assume that f is an expensive-to-compute black-box function. In this situation, optimization can only be attempted at a low number of function evaluations. The Efficient

*Corresponding author: Ecole Nationale Suprieure des Mines de Saint Etienne, Institut H. Fayol, 158, Cours Fauriel, 42023 Saint-Etienne cedex 2 - France
Tel : +33477426638
Email: hossein.mohammadi@emse.fr
†leriche@emse.fr
‡touboul@emse.fr

Global Optimization (EGO) algorithm [5, 6, 3] has become a standard for optimizing such expensive unconstrained continuous problems. Its efficiency stems from an embedded conditional Gaussian Process (GP, also known as kriging) which acts as a surrogate for the objective function.

The way the kriging model is learned from data points is essential to the EGO performance. A kriging model is mainly described by the associated kernel and this kernel determines the set of possible functions processed by the algorithm to make optimization decisions. Several methods alternative to cross-validation or Maximum Likelihood (ML) have been proposed to tune the kernel parameters. For example, a fully Bayesian approach is used in [2]. In [5], the process of estimating parameters and searching for the optimum are combined together through a likelihood which encompasses a targeted objective. In [10], the bounds on the parameter values are changing with the iterations following an a priori schedule. Nevertheless, we feel that the existing methods for learning kernel parameters are complex so that the basic phenomena taking place in the optimization when tuning the kernel cannot be clearly observed. This study allows to more deeply understand the influence of kriging parameters on the efficiency of EGO by studying the convergence of EGO with fixed parameters on a unimodal and a multimodal function. The effect of nugget is also investigated.

2 Kriging model summary

Let $\mathbf{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ be a set of n design points and $\mathbf{y} = \{f(\mathbf{x}^1), \ldots, f(\mathbf{x}^n)\}$ the associated function values at $\mathbf{X}$. Suppose the observations are a realization of a GP, $Y(\mathbf{x})$. The kriging model is the GP conditional on the observations, $Y(\mathbf{x}) \mid Y(\mathbf{x}^1) = \mathbf{y}_1, \ldots, Y(\mathbf{x}^n) = \mathbf{y}_n$, also written in a more compact notation, $Y(\mathbf{x}) \mid Y(\mathbf{X}) = \mathbf{y}$. The GP's prediction (kriging mean) and variance of prediction (kriging variance) at a point $\mathbf{x}$ are

$$m(\mathbf{x}) = \mu + \mathbf{r}(\mathbf{x})^\top \mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\mu), \tag{1}$$

$$s^2(\mathbf{x}) = \sigma^2 \left(1 - \mathbf{r}(\mathbf{x})^\top \mathbf{R}^{-1}\mathbf{r}(\mathbf{x})\right). \tag{2}$$

Here, μ and σ^2 are the process mean and variance, $\mathbf{1}$ is a $n \times 1$ vector of ones, $\mathbf{r}(\mathbf{x})$ is the vector of correlations between point $\mathbf{x}$ and the n sample points, $\mathbf{r}(\mathbf{x}) = [\mathrm{Cor}(Y(\mathbf{x}), Y(\mathbf{x}^1)), \ldots, \mathrm{Cor}(Y(\mathbf{x}), Y(\mathbf{x}^n))]$, and $\mathbf{R}$ is an $n \times n$ correlation matrix between sample points of general term $\mathbf{R}_{ij} = \mathrm{Cor}(Y(\mathbf{x}^i), Y(\mathbf{x}^j))$. The covariance function (i.e., the kernel) used here is the isotropic Matern 5/2 function defined as [8]

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \mathrm{Cor}(Y(\mathbf{x}), Y(\mathbf{x}')) = \sigma^2 \left(1 + \frac{\sqrt{5}\|\mathbf{x}-\mathbf{x}'\|}{\theta} + \frac{5\|\mathbf{x}-\mathbf{x}'\|^2}{3\theta^2}\right) \exp\left(-\frac{\sqrt{5}\|\mathbf{x}-\mathbf{x}'\|}{\theta}\right), \tag{3}$$

in which the parameter $\theta > 0$ is called *characteristic length-scale* and controls the correlation strength between pairs of response values. More generally, all stationary isotropic covariance functions have such a characteristic

length-scale. Anisotropic covariance functions have d such length-scales, one per dimension, as can be seen below with the usual tensor product kernel,

$$k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = \sigma^2 \prod_{i=1}^{d} k_i \left(x_i, x_i'; \theta_i\right) \tag{4}$$

In order to simplify the analysis, we will focus in the following on the isotropic case, $\theta_1 = \cdots = \theta_d = \theta$. The smaller θ, the least two response values at given points are correlated, and vice versa, see Fig. 1.

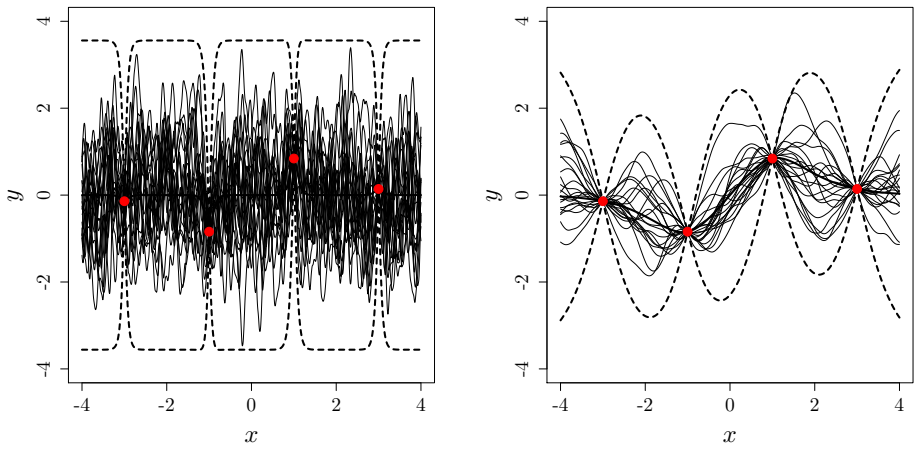

Figure 1: Kriging mean (thick solid line) along with the 95% confidence intervals (thick dashed lines), i.e., $m(\mathbf{x}) \pm 1.96 s(\mathbf{x})$, for $\theta = 0.1$ (left) and $\theta = 1$ (right). The thin lines are the sample paths of the GP. As θ changes, the class of possible functions considered for the optimization decision changes. Therefore, θ is a central decision for the optimization that deserves an in-depth study.

When a nugget, τ^2, is added to the model, the covariance function becomes

$$k_{\tau^2}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') + \tau^2 \delta(\mathbf{x}, \mathbf{x}'), \tag{5}$$

where $\delta(.,.)$ is the Kronecker's delta. Adding nugget to the model means that the observations are perturbed by an additive Gaussian noise $\mathcal{N}(0, \tau^2)$. The resulting kriging predictions, $m(\mathbf{x})$, are smoother as they no longer interpolate the observations[1]. Nugget also increases kriging variance throughout the search domain since, beside the changes in the covariance matrix $\mathbf{R}$, the term σ^2 becomes $\sigma^2 + \tau^2$ in Equation (2).

[1]Strictly speaking, if the covariance function of Eq. (5) is directly input into the kriging model, the trajectories are discontinuous and interpolating the observations. Therefore, often, nugget is only put on the covariance matrix and not on the covariance vector, which means that the observations are noisy but the prediction is not.

Classically here, the process mean and variance are estimated by the following ML closed-form expressions [8],

$$\hat{\mu} = \frac{\mathbf{1}^\top \mathbf{R}^{-1}\mathbf{y}}{\mathbf{1}^\top \mathbf{R}^{-1}\mathbf{1}} \quad , \quad \hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{1}\hat{\mu})^\top \mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu})}{n} \ , \tag{6}$$

so that the only kernel parameters left are θ and τ^2.

At any point $\mathbf{x}$ in $\mathcal{S}$, the improvement is defined as the random variable $I(\mathbf{x}) = \max(0, f_{min} - Y(\mathbf{x}) \mid Y(\mathbf{X}) = \mathbf{y})$ where f_{min} is the best objective function value observed so far. The improvement is the random excursion of the process at any point below the best observed function value. The expected improvement can be calculated analytically as

$$EI(\mathbf{x}) = \begin{cases} (f_{min} - m(\mathbf{x}))\Phi\left(\frac{f_{min}-m(\mathbf{x})}{s(\mathbf{x})}\right) + s(\mathbf{x})\phi\left(\frac{f_{min}-m(\mathbf{x})}{s(\mathbf{x})}\right) & \text{if } s(\mathbf{x}) > 0 \\ 0 & \text{if } s(\mathbf{x}) = 0 \ , \end{cases} \tag{7}$$

where Φ and ϕ denote the cumulative distribution function and probability density function of the standard normal distribution, respectively. $\mathrm{EI}(\mathbf{x})$ is positive everywhere in $\mathcal{S}$. If $\tau^2 = 0$, it is null at data points, but not when $\tau^2 > 0$. It is increasing when the kriging variance increases (at a fixed kriging mean) and when the kriging mean decreases (at a fixed kriging variance). The first term in Eq. (7) is dominated by the contribution of kriging mean to the improvement while the second term is dominated by the contribution of kriging variance. The EGO algorithm consists in the sequential maximization of EI, $\mathbf{x}^{n+1} = \arg\max_{x \in \mathcal{S}} EI(\mathbf{x})$ followed by the updating of the kriging model with $\mathbf{X} \cup \{\mathbf{x}^{n+1}\}$ and the associated responses $\mathbf{y}$.

3 EGO with fixed length-scale

We start by discussing the behavior of EGO with two different fixed length-scales (small and large). The magnitude of length-scale is measured with respect to the longest possible distance in the search space, $Dist_{max}$ which, in our d-dimensional search space is equal to $(UB - LB)\sqrt{d}$. θ is large if it is close to or larger than $Dist_{max}$ and vice versa. Here, $LB = -5$ and $UB = 5$. Fig. 5 illustrates the kriging models on the Ackley test function (defined below) in 1 dimension and the associated EIs for small and large length-scales.

3.1 Small characteristic length-scale

When θ is small, there is a low correlation between response values so that data points have an influence on the process only in their immediate neighborhood. As $\theta \to 0$ and away from the data points, the kriging mean and

variance of Equations (1) and (2) turn into the constants μ and σ^2, respectively, thus the EI becomes a constant flat function: when $\mathbf{x}$ is away from $\mathbf{x}^i$, $EI(\mathbf{x}) \to EI^{\text{asymp}} := (f_{min} - \hat{\mu})\Phi\left(\frac{f_{min}-\hat{\mu}}{\hat{\sigma}}\right) + \hat{\sigma}\phi\left(\frac{f_{min}-\hat{\mu}}{\hat{\sigma}}\right)$, where $\hat{\mu} \to \frac{\sum_{i=1}^{n} y^i}{n}$ and $\hat{\sigma}^2 \to \frac{\sum_{i=1}^{n}(y^i-\hat{\mu})^2}{n}$ since $\mathbf{R}$ tends to the identity matrix in Equation (6).

Proposition 1 (EGO iterates for small length-scale) *As the characteristic length-scale of the GP kernels tend to 0, the EGO iterates are located in a shrinking neighborhood of the most isolated best observed point.*

This proposition is explained and proved below.

Irrespectively of the function being optimized and the current DoE (provided the best observed point is uniquely defined), the set of design points created by EGO with small θ has characteristically repeated samples near the best observed points. An example is provided in Fig. 2 where $\theta = 0.001$. Elements of proof of this phenomenon is given below.

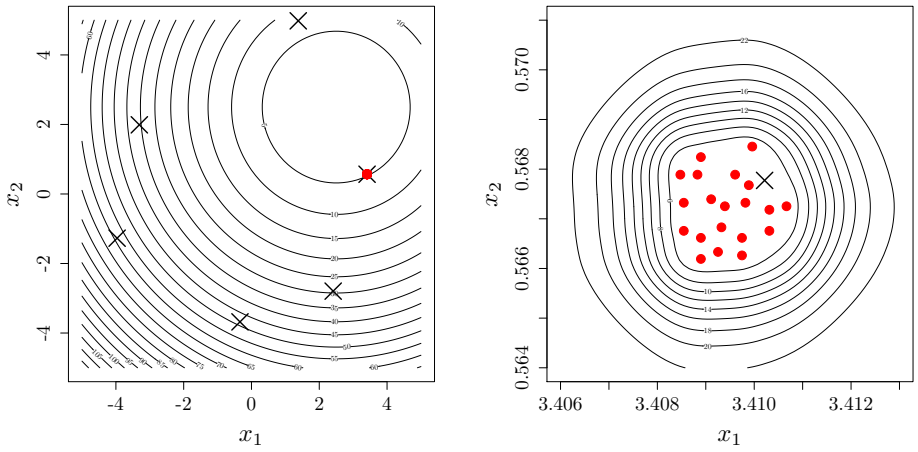

Figure 2: Left: search points obtained during 20 iterations of EGO with a small length-scale ($\theta = 0.001$) on the Sphere function whose contour lines are plotted. Crosses are the initial design points. The points accumulate in the vicinity of the design point with the lowest function value. Right picture: zoom around the best observed point; the contour lines show the kriging mean.

When the length-scale is small, the observations have a low range of influence. In the limit case, one can assume that in a vicinity of ith design point the correlation between $Y(\mathbf{x}^i)$ and the other observations is zero, i.e., $\text{Cor}(Y(\mathbf{x}^i), Y(\mathbf{x}^j)) \to 0$, $1 \leq j \leq n$, $j \neq i$, so that $R \to I$. Let $\mathbf{x}$ be in the neighborhood of $\mathbf{x}^i$, $B_\epsilon(\mathbf{x}^i) = \left\{\mathbf{x} \in \mathcal{S} : \ \|\mathbf{x} - \mathbf{x}^i\| \leq \epsilon\right\}$, for a sufficiently small ϵ and away from the other points of the Design of Experiments (DoE)

$j \neq i$ so that the correlation vector tends to $\mathbf{r}(\mathbf{x}) \to [0, \ldots, 0, r, 0, \ldots 0]$ where $r = \mathrm{Cor}(Y(\mathbf{x}), Y(\mathbf{x}^i))$. In this situation, the kriging mean and variance can be fully expressed in terms of the correlation r (a scalar in $[0, 1]$):

$$m(r) = \hat{\mu} + r(y_i - \hat{\mu}) = \hat{\mu}(1 - r) + r y_i, \tag{8}$$

$$s^2(r) = \hat{\sigma}^2(1 - r^2), \tag{9}$$

It is visible from the above equations that, among the points of the DoE, the expected improvement will be the largest near the best observed point as, for any given r, the variance will be the same and the mean will be the lowest. If many points of the DoE share the same best performance f_{min}, we will consider $\mathbf{x}^{min}$, the most isolated[2] one. By setting $y_i = f_{min}$ in Eqs. (8) and (9), the expected improvement (Eq. (7)) in the vicinity of the best observed point becomes,

$$\begin{aligned} EI(r) \; = & (1 - r)(f_{min} - \hat{\mu})\Phi\left(\frac{f_{min} - \hat{\mu}}{\hat{\sigma}}\sqrt{\frac{1 - r}{1 + r}}\right) + \\ & \hat{\sigma}\sqrt{1 - r^2}\phi\left(\frac{f_{min} - \hat{\mu}}{\hat{\sigma}}\sqrt{\frac{1 - r}{1 + r}}\right). \end{aligned} \tag{10}$$

Dividing both sides of Equation (10) by $\hat{\sigma}$ and introducing the new variable $A := \frac{f_{min} - \hat{\mu}}{\hat{\sigma}}$, the normalized expected improvement $EI(r)/\hat{\sigma}$, reads

$$EI(r)/\hat{\sigma} = (1 - r)A\Phi\left(A\sqrt{\frac{1 - r}{1 + r}}\right) + \sqrt{1 - r^2}\phi\left(A\sqrt{\frac{1 - r}{1 + r}}\right). \tag{11}$$

The normalized improvement is handy in that, for small length scale, it sums up what happens for all objective functions, design of experiments and kernels in terms of only two scalars, the correlation r and A. Note that because $f_{min} \leq y_i \;,\; \forall i$, $A \leq 0$. Instances of normalized EI are plotted for a set of A's in $[-2, -0.001]$ in the left of Fig. 3. The value of EI when $r \to 0^+$ is the asymptotic value of expected improvement as $\mathbf{x}$ moves away from data points. The maximum of EI (equivalently $EI/\hat{\sigma}$) is reached at $r^\star$ which is strictly larger than 0. All the values of $r^\star$ are represented as a function of A in the right plot of Fig. 3. As A decreases (i.e., f_{min} further drops below $\hat{\mu}$, or the best observation improves with respect to the other observations), $r^\star$ tends to 1, that is EGO will create the next iterate closer to $\mathbf{x}^{min}$, which makes sense since the point gets better. Vice versa, as the advantage of the best observation reduces (A diminishes), $r^\star$ approaches 0, which means that EGO will put the next iterate further from $\mathbf{x}^{min}$. Note that the analytical formulas for the first and second derivative of EI with respect to r are given in Appendix A.

[2] the most isolated in terms of the metric used by the covariance functions of the GP.

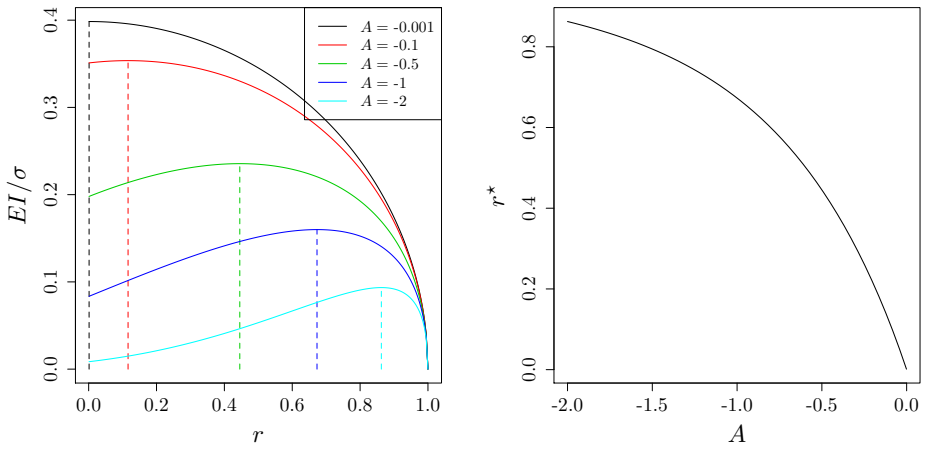

Figure 3: Left: Normalized EI as a function of $r \in]0, 1]$ in the vicinity of the sample point with the lowest function value for a small length-scale. Right: location of the next EGO iterate ($r^\star$ where EI is maximized) as a function of A.

3.2 Large characteristic length-scale

Proposition 2 (EGO iterates for large length-scale) *As the characteristic length-scale of the GP kernels increases, $\theta \to \infty$, the EGO algorithm degenerates into the sequential minimization of the kriging mean $m(\boldsymbol{x})$.*

This behavior of EGO can be understood by seeing that as the length-scale increases, the points have more influence on each other and the uncertainty, as described by kriging variance $s^2(\mathbf{x})$ in Equation (2), vanishes. Then, we will see that maximizing the expected improvement is equivalent to minimizing the kriging mean when kriging variance is null.

Let us demonstrate the above statements. We first establish that the term $\mathbf{r}(\mathbf{x})^\top \mathbf{R}^{-1} \mathbf{r}(\mathbf{x})$ in the kriging variance of Equation (2) tends to 1. As $\theta \to \infty$, all the responses $Y(x)$ are strongly correlated, therefore $\mathbf{r}(\mathbf{x})$ and $\mathbf{R}$ become a vector and a matrix of 1's. This matrix $\mathbf{R}$ has only one non-zero eigenvalue that equals n, the matrix size [1]. The corresponding eigenvector is $\mathbf{v} = \frac{\sqrt{n}}{n}(1, \ldots, 1)^\top$. To invert such a non-invertible matrix, we use *Moore-Penrose pseudoinverse* [9], which is equivalent to regularizing it with a very small nugget (see [7]). The pseudoinverse of $\mathbf{R}$, denoted by $\mathbf{R}^\dagger$, is

$$\mathbf{R}^\dagger = [\mathbf{v}\ \mathbf{W}] \begin{bmatrix} \frac{1}{n} & \mathbf{0}_{1\times(n-1)} \\ \mathbf{0}_{(n-1)\times 1} & \mathbf{0}_{(n-1)\times(n-1)} \end{bmatrix} [\mathbf{v}\ \mathbf{W}]^\top, \tag{12}$$

in which $\mathbf{W}$ contains the $n-1$ eigenvectors associated with the zero eigenvalues. Regularizing $\mathbf{R}^{-1}$ as $\mathbf{R}^\dagger$ in $\mathbf{r}(\mathbf{x})^\top \mathbf{R}^{-1} \mathbf{r}(\mathbf{x})$ and since $\mathbf{r}(\mathbf{x})^\top \to (1, \ldots, 1)$

as $\theta \to \infty$, it is easy to show that $\mathbf{r}(\mathbf{x})^\top \mathbf{R}^\dagger \mathbf{r}(\mathbf{x}) = 1$. As a result, $s^2(\mathbf{x}) \to 0$ and $EI(\mathbf{x}) \to f_{min} - m(\mathbf{x})$. In this case, the EGO search degenerates to an iterative minimization and updating of the kriging mean $m(\mathbf{x})$.

Minimizing kriging mean does not define a valid global optimization scheme for two reasons. Firstly, because premature convergence occurs as soon as the minimum of $m(\mathbf{x})$ coincides with an observation of the true function [5]: when $m(\mathbf{x}^{n+1}) = f(\mathbf{x}^{n+1})$ where $\mathbf{x}^{n+1} = \arg\min_{\mathbf{x}\in\mathcal{S}} m(\mathbf{x})$, the EGO iterations with large θ stop producing new points, however $\mathbf{x}^{n+1} \cup \mathbf{X}$ may not even contain a local optimum of f. Secondly, it should be remembered that the kriging mean discussed here is that stemming from large length-scale, which may not allow an accurate prediction of the objective function considered: it would suit a function like the sphere with a Matérn kernel, but it would not suit a multimodal function like Ackley.

The DoE created by EGO with large θ can vary greatly depending on the function and the initial DoE. On the one hand, if the function is regular and well predicted by $m()$ around $\mathbf{x}^{n+1}$, like the Sphere function, the kriging mean rapidly converges to the true function and points are accumulated in this region which may or not be the global optimum. Fig. 4 illustrates both situations (true and false convergence) with the DoEs created by an EGO algorithm with large length-scale on a unimodal and a multimodal function (Sphere and Rastrigin functions, respectively). The Rastrigin function is defined as

$$f_{\text{Rastrigin}}(\mathbf{x}) = 10d + \sum_{i=1}^{d} \left(x_i^2 - 10\cos(2\pi x_i)\right). \tag{13}$$

On the other hand, if $m(\mathbf{x}^{n+1})$ is different from $f(\mathbf{x}^{n+1})$, the kriging mean changes a lot between iterations because new observations have a long range influence. The kriging mean overshoots observations in both upper and lower directions (cf. the dotted blue curve in the upper left plot of Fig. 5). The resulting DoE is more space-filling than the DoE of small length scale. An example of such DoE is provided at the bottom right of Fig. 5.

3.3 Comparison of EGO with fixed and adapted length-scale

In the sequel, the efficiency of EGO with different fixed length-scale is compared with the standard EGO whose length-scale is learned by ML. Tests are carried out on two isotropic functions, the unimodal sphere and the highly

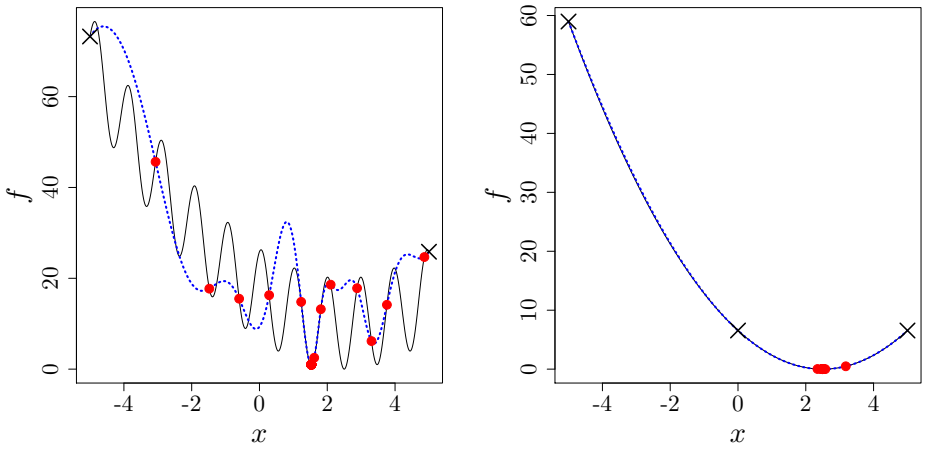

Figure 4: DoE created by EGO with $\theta = 100$. For such a large θ, the global search turns into the sequential minimization of the kriging mean. Left: premature convergence of the algorithm in a local minimum of the Rastrigin function because $m(\mathbf{x}^{n+1}) = f(\mathbf{x}^{n+1})$. The true optimum is at $x^\star = 2.5$ in the neighboring basin of attraction. Right: the algorithm converges to the global minimum of the unimodal Sphere function. In both functions the global minimum is located at 2.5.

multimodal Ackley functions:

$$f_{\text{Sphere}}(\mathbf{x}) = \sum_{i=1}^{d} (x_i)^2, \tag{14}$$

$$f_{\text{Ackley}}(\mathbf{x}) = -20 \exp\left(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d} \cos(2\pi x_i)\right) + 20 - \exp(1). \tag{15}$$

Each optimization is repeated 5 times on 5 dimensional instances of the problems, $d = 5$. The initial DoE is fixed and has size $3 \times d$. The search length is $70 \times d$. To allow comparisons of the results, the functions are scaled (multiplied) by $\frac{2}{f_{DoE}^{max} - f_{DoE}^{min}}$, where f_{DoE}^{min} and f_{DoE}^{max} are the smallest and the largest value of function f in the initial DoE.

Fig. 6 shows the results of the comparison in terms of median objective functions. Moreover, the first and the third quartiles are plotted in Fig. 7. The θ values belong to the set $\{0.01, 0.1, 1, 5, 10, 20\}$. On both test functions, the algorithm does not converge quickly towards the minimum when $\theta = 0.01$ or $\theta = 0.1$ because, as explained in Section 3, it focuses on the neighborhoods of the best points found early in the search. On the Sphere function, EGOs with large length-scale, $\theta = 20$ or $\theta = 10$, have performances equivalent to that of the standard EGO. Indeed, the Sphere function is very smooth and, as can be seen on the rightmost plot of Fig. 6,

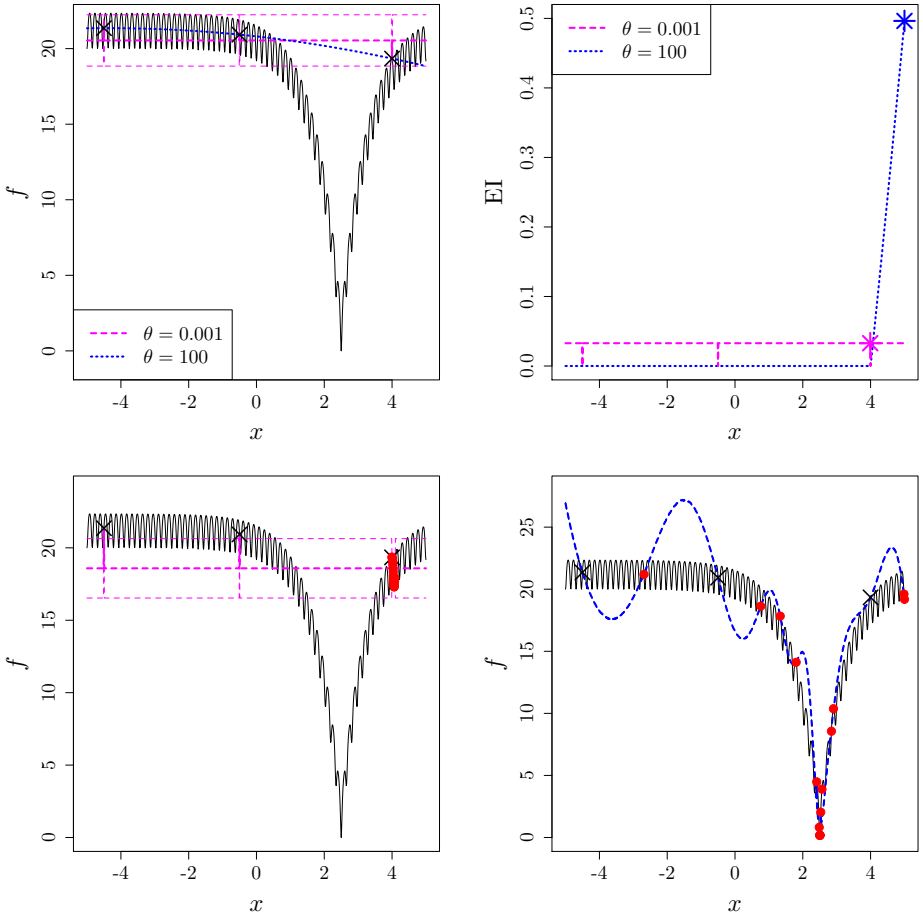

Figure 5: Ackley function (black solid line and defined in (15)) approximated by a kriging model (mean ± std. deviation, thick/thin lines) with $\theta = 0.001$ (dashed pink) and $\theta = 100$ (dotted blue). The crosses are the initial DoE. Top, right: EIs at iteration 1 with the stars indicating the EI maximums. Bottom, red bullets: DoEs created by EGO after 20 iterations with $\theta = 0.001$ (left) and $\theta = 100$ (right).

ML estimates of θ are equal to 20 (the upper bound of the ML) rapidly after a few iterations. With the multimodal Ackley function, the best fixed θ is equal to 1. It temporarily outperforms the standard EGO at the beginning of the search (until about 70 evaluations) but then ML allows decreasing the θ's until about 0.5 (see rightmost plot) and fine tuning the search in the already located high performance region. Note however that this early advantage of $\theta = 1$ over the adapted θ seem to be dependent on the initial DoE (cf. experiment with an alternative DoE in Fig. 8).

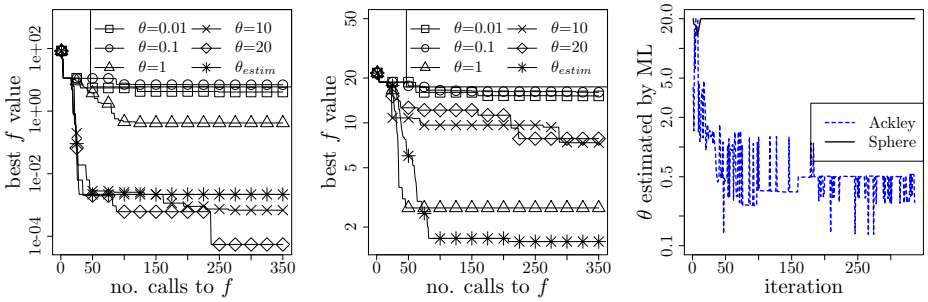

Figure 6: Median of the best objective function vs. number of calls of standard EGO and EGO with different fixed length-scale on the Sphere (left) and the Ackley (middle) functions, $d = 5$. Right: evolution of θ learned by ML in standard EGO.

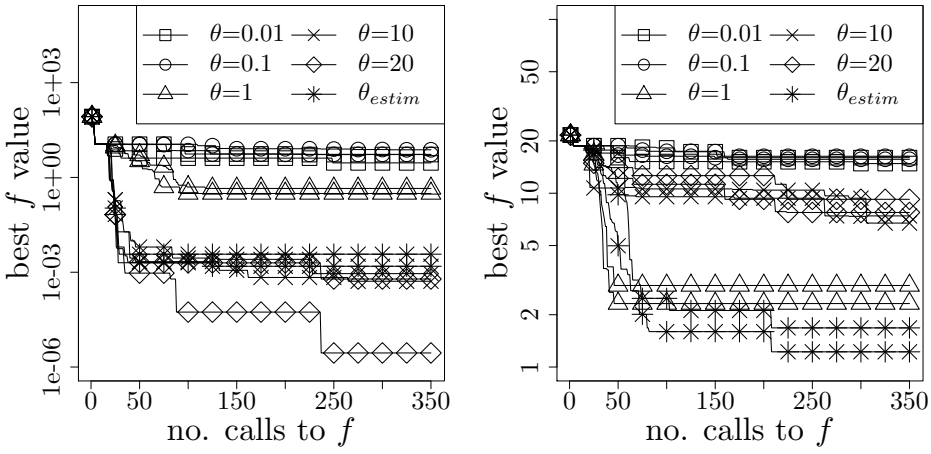

Figure 7: Dispersion of the results of Fig. 6 : first and the third quartiles of the results for the Sphere (left) and Ackley (right) functions.

In order to investigate the effect of initial DoE on the above results, we repeat the same experiments with another fixed DoE. The results with the new DoE are given in Fig. 8. These results are similar to those already

reported in Fig. 6, therefore showing a low sensitivity of EGO to the initial DoE. The main difference is visible in the initial iterations (before 100 calls) for the multimodal Ackley function and questions the early advantage at using $\theta = 1$ over θ adapted by ML.

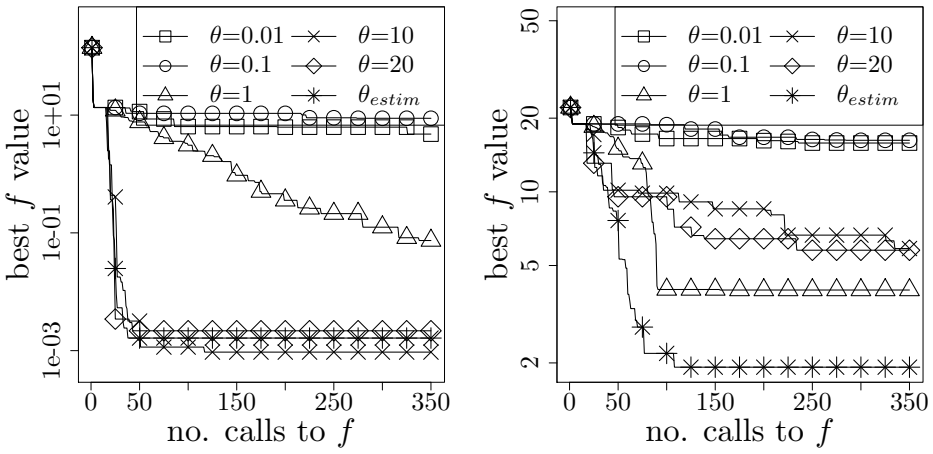

Figure 8: Median of the best objective function vs. number of calls of standard EGO and EGO with different fixed length-scale on the Sphere (left) and the Ackley (middle) functions, $d = 5$. Although the initial DoE is different from the one used in Fig. 6, the EGO performance does not change a lot.

A complementary view on convergence, focusing on distances to the optimum in the x-space and the whole set of search points created, as opposed to the objective function of the best point in the convergence plots (e.g., Fig. 6), is given in Fig. 9. Each curve represents the density of search points closer to the global minimum than a given distance. The procedure for calculating this density is to divide the number of points closer to the global minimum by the total number of the points of the search (here 350 when $d = 5$). The distances are normalized by dividing them by the square root of the problem dimension. For small distances to the optimum ($< 0.3 \times \sqrt{d}$), the algorithms hierarchy recovered from these graphs is based on the best points and is similar to that of Fig. 6. For larger distances, we find out that EGO with fixed $\theta = 1$ performs very well at creating many points within a distance of $1 \times \sqrt{d}$ to the optimum.

4 Effect of nugget on EGO convergence

To investigate the effect of nugget on EGO, we carry out the same test protocol as above but the length-scales are set by ML and two scenarios are considered: 1) the nugget τ^2 is estimated by ML, 2) a fixed nugget is taken

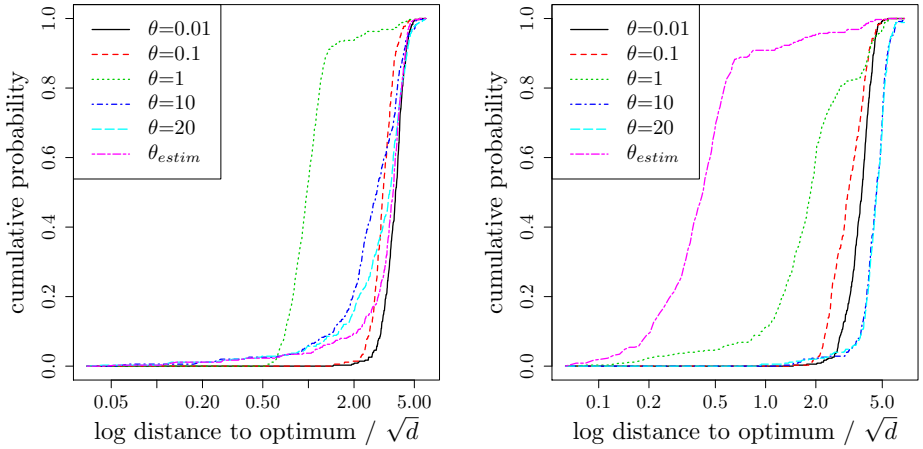

Figure 9: Density of points closer to the optimum than a given distance on Sphere (left) and Ackley (right) functions. Each curve is the median of 5 runs.

from the set $\tau^2 \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, 0\}$ ($\tau^2 = 0$ means no nugget). Fig. 10 shows the results. For both test functions, when the nugget value is large (10^{-2} or 10^{-4} or ML estimated on Ackley), EGO exhibits the worst performances: it does not converge faster and stops further from the optimum. The reason is that a large nugget deteriorates the interpolation quality of a kriging model when observations are not noisy like here. On the Sphere function, EGO rapidly locates the area of the optimum but the EI without nugget, which is null at data points, pushes the search away from it. However, a nugget value equal to 10^{-6} or 10^{-8} hardly slows down convergence and significantly improves the accuracy with which the optimum is found. Indeed, by increasing the uncertainty $s^2(\mathbf{x})$ everywhere including in the immediate vicinity of data points, where it would be null without nugget, nugget increases the EI there and allows a higher concentration of EGO iterates near the best observed point. The nugget learned by ML on the Sphere tends to 0 which, as just explained, is not the best setting for optimization.

On Ackley, besides large nugget values ($\tau^2 \geq 10^{-4}$) which significantly degrade the EGO search, values ranging from $\tau^2 = 0$ to 10^{-6} do not notably affect performance. In this case, the global optimum is not accurately located after $70 \times d$ evaluations of f, there is no need to allow through nugget an accumulation of points near the best observation.

Note that on both functions, when considering the best point found so far, ML estimation of nugget is not a good strategy. Finally, the dispersion of all the search points the across the x-space is characterized in Fig. 11 through the number (the density) of points closer to the optimum than a

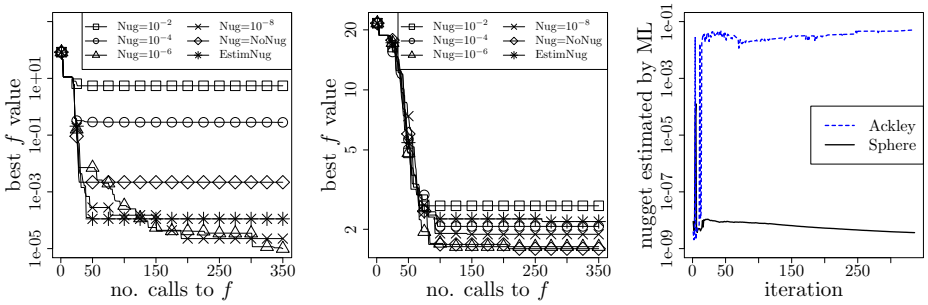

Figure 10: Median of the best objective function vs. number of calls to f for EGO with different nugget values on the Sphere (left) and Ackley (middle) functions in dimension 5. Right: ML estimated nugget, τ^2, vs. number of calls to f.

given distance (cf. previous section for a more detailed definition). For the Sphere function, $\tau^2 = 10^{-6}, 10^{-4}$ and 10^{-2} allow locating more points in a larger neighborhood of the optimum, respectively. For the Ackley function, no to moderate ($\tau^2 = 10^{-4}$) nuggets produce similar densities of points around the optimum; $\tau^2 = 10^{-2}$ seems to be often missing high performance areas; the ML estimate of τ^2, which after initial oscillations between 0 and 5.10^{-2}, stabilizes over 5.10^{-2}, puts 7% of the search points within a distance of $0.07\times\sqrt{d}$ of the optimum (which makes it the best strategy at this distance to the optimum) but then puts the remaining points far from the optimum.

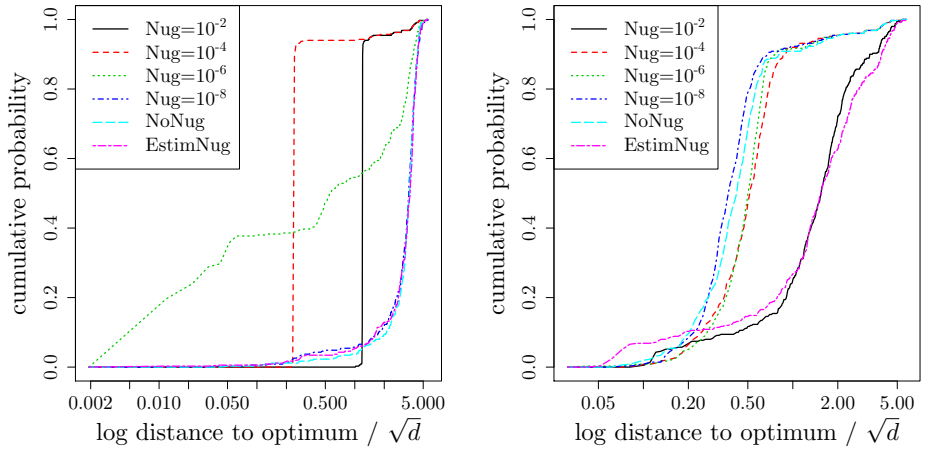

Figure 11: Cumulative probability of search points under different scenarios of nugget values on Sphere (left) and Ackley (right) function.

5 Concluding remarks

To sum up, this paper carefully explains the DoEs generated by EGO with fixed length-scale and nugget. In terms of performance, ML estimation of the length-scale is a good choice but ML estimation of nugget is not recommended (a fixed small nugget value should be preferred). As a perspective, EGO strategies starting with a large fixed length-scale and then decreasing it while keeping a small amount of nugget should be efficient while avoiding ML estimations which require $O(n^3)$ computations [4].

Acknowledgments.

The authors would like to acknowledge support by the French national research agency (ANR) within the Modèles Numériques project "NUMerical Black-Box Optimizers" (NumBBO).

A Expected Improvement and its derivatives for small length-scale

When the length-scale is small, the normalized expected improvement tends to the following analytical expression

$$\frac{EI(r)}{\sigma} = (1-r)A\Phi\left(A\sqrt{\frac{1-r}{1+r}}\right) + \sqrt{1-r^2}\phi\left(A\sqrt{\frac{1-r}{1+r}}\right), \tag{16}$$

where r is the correlation with the best observed point and $A = \frac{f_{min}-\hat{\mu}}{\hat{\sigma}}$. Such expression applies to any objective functions, designs of experiment and kernels as long as the length-scale tends to 0. We want to calculate the first and the second derivatives of the normalized expected improvement with respect to r: To do so, we need to calculate the derivative of each term. Here, we present the derivatives of the terms $\Phi\left(A\sqrt{\frac{1-r}{1+r}}\right)$, $\phi\left(A\sqrt{\frac{1-r}{1+r}}\right)$ and $\sqrt{\frac{1-r}{1+r}}$ which are

$$\frac{\partial}{\partial r}\Phi\left(A\sqrt{\frac{1-r}{1+r}}\right) = A\left(\frac{\partial}{\partial r}\sqrt{\frac{1-r}{1+r}}\right)\phi\left(A\sqrt{\frac{1-r}{1+r}}\right), \tag{17}$$

$$\frac{\partial}{\partial r}\phi\left(A\sqrt{\frac{1-r}{1+r}}\right) = -\left(A\sqrt{\frac{1-r}{1+r}}\right)\frac{\partial}{\partial r}\left(A\sqrt{\frac{1-r}{1+r}}\right)\phi\left(A\sqrt{\frac{1-r}{1+r}}\right), \tag{18}$$

$$\frac{\partial}{\partial r}\sqrt{\frac{1-r}{1+r}} = \frac{-\sqrt{1-r}}{2(1+r)^{3/2}} - \frac{1}{2\sqrt{1-r^2}}. \tag{19}$$

After calculating all the derivatives and simplification, the first derivative of $\frac{EI(r)}{\sigma}$ with respect to r can be written as

$$\frac{\partial EI(r)}{\sigma \partial r} = -A\Phi\left(A\sqrt{\frac{1-r}{1+r}}\right) - \frac{r}{\sqrt{1-r^2}}\phi\left(A\sqrt{\frac{1-r}{1+r}}\right). \tag{20}$$

In Fig. 12, the first derivative of $\frac{EI(r)}{\sigma}$ for different values of A is numerically calculated. The location of a stationary point, $r^\star$, is where $\frac{\partial EI(r^\star)}{\sigma \partial r} = 0$, and it is also numerically estimated.

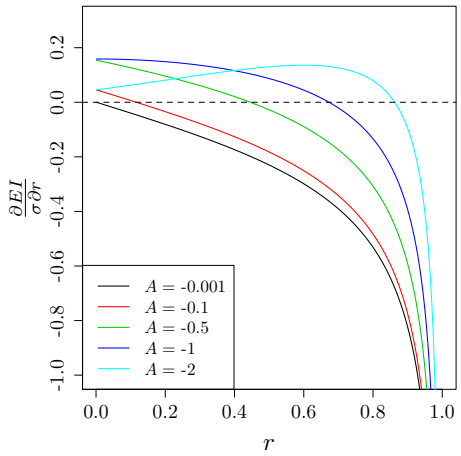

Figure 12: First derivative of $\frac{EI(r)}{\sigma}$ with respect to r for different values of A. The location of the stationary point becomes closer to $r = 0$ as $A \to 0^-$. In other words, for (negative) values of A different from 0, r is finite and the maximum of the EI is achieved near the best known point.

To determine the nature of the stationary points, the second derivative of $\frac{EI(r)}{\sigma}$, i.e., $\frac{\partial^2 EI}{\sigma \partial r^2}$, is required which is:

$$\frac{\partial^2 EI}{\sigma \partial r^2} = \left[\frac{A^2(1-r) - (1+r)}{(1+r)^{5/2}(1-r)^{3/2}}\right]\phi\left(A\sqrt{\frac{1-r}{1+r}}\right). \tag{21}$$

In the left picture of Fig. 13 the second derivative of $\frac{EI(r)}{\sigma}$, $\frac{\partial^2 EI}{\sigma \partial r^2}$, with the same A values as used in Fig. 12 is shown. In the right picture, the value of $\frac{\partial^2 EI}{\sigma \partial r^2}$ is plotted at the stationary points $r^\star$. It can be seen that the second derivatives are always negative. In other words, the curvature of the function $\frac{EI(r)}{\sigma}$ at any stationary points is negative and the function has a maximum there.

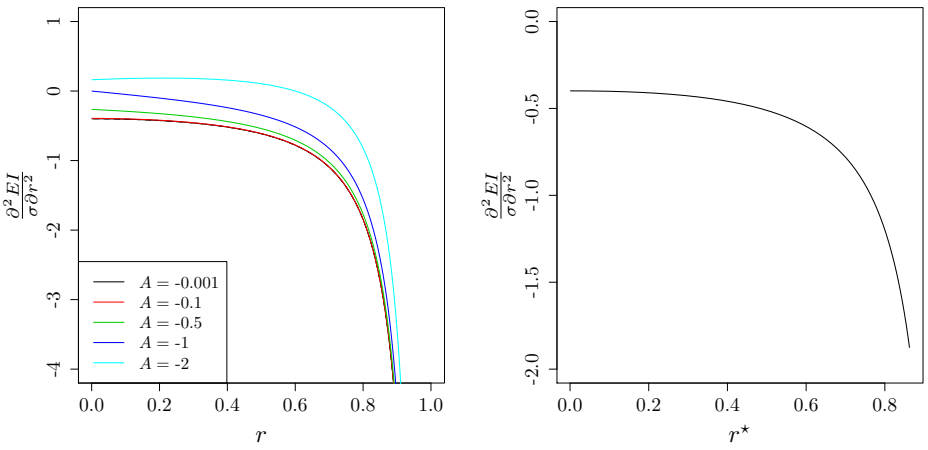

Figure 13: Left: second derivative of $\frac{EI(r)}{\sigma}$ when A equals to $-2, -1, -0.5, -0.1, -0.01$. The second derivative is negative most of the time excepted when A is small and r is close to 0 (compare to Fig. 3). Right: the value of $\frac{\partial^2 EI}{\sigma \partial r^2}$ is plotted for different values of $r^\star$. This curvature is always negative.

References

[1] Andrianakis, I., Challenor, P.G.: The effect of the nugget on Gaussian process emulators of computer models. Computational Statistics & Data Analysis 56(12), 4215–4228 (2012)

[2] Benassi, R., Bect, J., Vazquez, E.: Robust gaussian process-based global optimization using a fully bayesian expected improvement criterion. In: Coello, C. (ed.) Learning and Intelligent Optimization, Lecture Notes in Computer Science, vol. 6683, pp. 176–190. Springer Berlin Heidelberg (2011)

[3] Brochu, E., Cora, V.M., de Freitas, N.: A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Tech. Rep. TR-2009-23, Department of Computer Science, University of British Columbia (November 2009)

[4] Cressie, N., Johannesson, G.: Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(1), 209–226 (2008)

[5] Jones, D.R.: A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization 21, 345–383 (2001)

[6] Kleijnen, J.P.: Kriging metamodeling in simulation: A review. European Journal of Operational Research 192(3), 707 – 716 (2009), http://www.sciencedirect.com/science/article/pii/S0377221707010090

[7] Mohammadi, H., Le Riche, R., Durrande, N., Touboul, E., Bay, X.: An analytic comparison of regularization methods for Gaussian Processes. Research report, Ecole Nationale Supérieure des Mines de Saint-Etienne ; LIMOS (Jan 2016), https://hal.archives-ouvertes.fr/hal-01264192

[8] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning, The MIT Press (2005)

[9] Strang, G.: Linear Algebra and Its Applications. Brooks Cole (1988)

[10] Wang, Z., Zoghi, M., Hutter, F., Matheson, D., de Freitas, N.: Bayesian optimization in high dimensions via random embeddings. In: International Joint Conferences on Artificial Intelligence (IJCAI) (2013)